UNIVERSIDADE FEDERAL DE ALFENAS - UNIFAL-MG

EDUARDA DE MELLO GUARNIERI

ALÉM DO GOL: UMA BREVE REVISÃO DE LITERATURA SOBRE MODELOS DE REGRESSÃO LINEAR COM APLICAÇÃO DE UM MODELO DE PROBABILIDADE LINEAR PARA ANALISAR A PROBABILIDADE DE VITÓRIA DO PALMEIRAS EM UM JOGO DO CAMPEONATO BRASILEIRO DE 2023

VARGINHA-MG
2023

EDUARDA DE MELLO GUARNIERI

ALÉM DO GOL: UMA BREVE REVISÃO DE LITERATURA SOBRE MODELOS DE REGRESSÃO LINEAR COM APLICAÇÃO DE UM MODELO DE PROBABILIDADE LINEAR PARA ANALISAR A PROBABILIDADE DE VITÓRIA DO PALMEIRAS EM UM JOGO DO CAMPEONATO BRASILEIRO DE 2023

Trabalho de conclusão de Piepex apresentado ao Instituto de Ciências Sociais Aplicadas da Universidade Federal de Alfenas como requisito parcial à obtenção do título de Bacharel em Ciência e Economia.

Orientador: Prof. Dr. Manoel Vítor de Souza Veloso

Coorientadora: Prof.ª Dr.ª Letícia Lima Milani Rodrigues

VARGINHA- MG
2023

RESUMO

A aplicação de análises estatísticas no futebol tem crescido nos últimos anos. O uso de métodos como modelos de regressão se mostra eficiente, pois realiza a previsão de determinada característica, influenciada por variáveis como número de gols, finalizações, aproveitamento de cada jogador, dentre outros, que possam influenciar na vitória de um time. Nesse sentido, o objetivo deste trabalho consiste em uma revisão de literatura sobre modelo de regressão linear, com a aplicação de um modelo de probabilidade linear para explicar a probabilidade de vitória do Palmeiras em um jogo do Campeonato Brasileiro de 2023, variável *dummy* que recebe 1 para vitória e 0 caso ela não ocorra. Para isso, foram coletados dados sobre 4 variáveis que influenciam nesse resultado, sendo elas Mando de jogo, Número de finalizações por jogo, Porcentagem de posse de bola por jogo e Número de gols por jogo, construindo e ajustando, por meio do método backward, um MPL para analisar a probabilidade de vitória. Os resultados encontrados mostraram que o modelo consegue explicar, aproximadamente, 48,53% da variabilidade total das vitórias a um nível de 5% de significância, e os resíduos seguem os pressupostos de independência e homocedasticidade, mas é considerada uma normalidade assimptótica a eles.

Palavras-chave: futebol; previsão; regressão; vitória.

ABSTRACT

The application of statistical analysis in football has grown in recent years. The use of methods such as regression models has proven to be efficient, as it predicts a specific outcome influenced by variables such as the number of goals, shots on target, player performance, among others, which may impact a team's victory. In this context, the objective of this study is to conduct a literature review on linear regression models, with the application of a linear probability model to explain the probability of Palmeiras winning a match in the 2023 Brazilian Championship. The dependent variable is a dummy variable that receives a value of 1 for a win and 0 otherwise. To achieve this, data on four variables influencing this outcome were collected: Home/Away, Number of shots per game, Ball possession percentage per game, and Number of goals per game. A logistic regression model was constructed and adjusted using the backward method to analyze the probability of victory. The results indicate that the model can explain approximately 48.53% of the total variability in victories at a 5% significance level. Additionally, the residuals meet the assumptions of independence and homoscedasticity, but asymptotic normality is assumed for them.

Keywords: football; prediction; regression; victory.

LISTA DE TABELAS

SUMÁRIO

1-INTRODUÇÃO

O futebol, esporte mais amado do mundo, tem ganhado cada vez mais estudos, em especial na área estatística, para que um time consiga vantagens sobre os times concorrentes, uma vez que as partidas estão sendo decididas por detalhes. No mundo atual, os times estão cada vez mais equilibrados, e qualquer melhoria que possa ser feita em parâmetros como finalizações, porcentagem de posse de bola, tipos de faltas marcadas, aproveitamento de cada posição, dentre outros, é buscada pelos treinadores (Aoki, 2010).

Com isso, para mostrar essa relação de vitória sobre o outro time, é inferido que algumas variáveis afetam esse resultado. Nesse sentido, a aplicação de um modelo de regressão, seja ele simples, múltiplo, linear, generalizado, ou outros, se mostra como uma ferramenta efetiva para analisar as influências das variáveis escolhidas na vitória de um time de futebol.

A regressão linear nada mais é do que "um instrumento estatístico para, simplesmente, resumir dados e informações" (Chein, 2019). O modelo geral de regressão linear é aquele que diz respeito à relação linear da variável dependente **Y** com a variação nas variáveis independentes **X.** Além disso, todo modelo possui uma parcela de erro, também chamado de resíduo, que é a diferença entre o valor observado e o valor estimado pelo modelo, parcela que deve representar uma porcentagem baixa da variabilidade total do modelo. Para isso, o método dos mínimos quadrados ordinários, o MQO, minimiza a soma dos quadrados dos erros, maximizando a soma dos quadrados da regressão, apresentando maior participação na explicação do modelo. Ainda, nem todas as variáveis escolhidas podem ser significativas no modelo, sendo necessário realizar o ajuste por meio de métodos como *backward, forward* e *stepwise*, com a análise do AIC de cada modelo ajustado e, para encerrar a análise, é feita uma verificação de pressupostos sobre esses resíduos, como independência, normalidade e homocedasticidade.

Dessa forma, este trabalho buscou realizar uma breve revisão de literatura sobre modelos de regressão linear, com a aplicação de um modelo de probabilidade linear para analisar a probabilidade de vitória do Palmeiras em um jogo do Campeonato Brasileiro de 2023. A escolha deste tipo de modelo se deu pela possibilidade de inclusão de variáveis *dummy*, que são variáveis categóricas binárias que recebem probabilidade 0 ou 1 conforme a categoria escolhida. Para isso, foram utilizados dados do site FootStats acerca de 5 informações: Vitória no jogo, que é a variável independente *dummy*, recebendo 0 se não ocorre vitória, e 1 se ocorre,

Mando de jogo, Número de finalizações por jogo, Porcentagem de posse de bola por jogo e Número de gols por jogo. Essas 4 últimas variáveis foram consideradas como possíveis influências na vitória do time em uma partida, construindo um modelo completo, ajustado pelo método *backward* de seleção de variáveis independentes, por meio do software estatístico R.

Este trabalho está organizado em cinco seções, incluindo a introdução e as considerações finais. A segunda seção apresenta uma revisão de literatura acerca de modelos de regressão e suas aplicações. A terceira seção aborda a fundamentação teórica necessária para uma análise de um modelo de regressão, bem como o método dos mínimos quadrados ordinários, métodos de seleção de variáveis independentes, os pressupostos para os resíduos e características de um modelo de probabilidade linear (MPL). A quarta seção trata de uma aplicação de um MPL para a previsão da probabilidade de vitória do Palmeiras em uma partida do Campeonato Brasileiro de 2023. Por fim, são feitas as considerações finais do estudo.

2-REVISÃO DE LITERATURA

O termo regressão foi proposto pela primeira vez por Francis Galton em 1885, enquanto realizava um estudo sobre as alturas de pais e filhos. Em seus estudos, observou que a altura observada nos pais não era transmitida completamente aos filhos, em que a altura deles progredia para um ponto médio da população, e a regressão foi tomada como o termo para definir a observação dessa tendência (Santos, 2016). Desde então, os modelos de regressão vêm sendo aplicados em diversas áreas do conhecimento, com enfoque para áreas das ciências sociais aplicadas e da saúde, sendo utilizados na metodologia de diversos trabalhos publicados que exploram os modelos e os ajustam (Rodrigues, 2012).

No mundo globalizado atual, cada vez mais as empresas buscam melhorar sua rentabilidade, diminuindo custos e otimizando suas atividades. Nesse cenário, os métodos estatísticos são utilizados como ferramentas de suporte para as tomadas de decisões gerenciais, como modelos de calcular a esperança sobre determinado valor ou produto, ou seja, uma previsão (Bastos; Guimarães; Severo, 2015).

Como uma extensão dos modelos lineares, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLG), pertencendo a distribuição da variável resposta à uma família exponencial. Além disso, para Cavalcante e Reinaldo (2019), os MLG's têm apresentado uma demanda crescente, pois são modelos flexíveis na variável resposta, na precificação

multivariada e na capacidade de inferência. De acordo com David (2021), para a estimação dos parâmetros da MLG, “podem ser utilizados vários métodos de estimação como qui-quadrado mínimo, o Bayesiano e o de máxima verossimilhança”.

Um MLG possui alguns componentes básicos: o componente aleatório, o componente sistemático e a função de ligação. O primeiro diz respeito às variáveis aleatórias Yi dependentes, enquanto o segundo estabelece a forma com que as variáveis explicativas Xi compõem o modelo, e o terceiro é o elemento que liga os dois anteriores (David, 2021).

No ramo previdenciário, no trabalho desenvolvido por Pinto et al (2019), considera como variáveis independentes o PIB, trabalho informal, taxa de desemprego, salários médios, taxa de inatividade, taxa de inflação e juros, para Brasil, Canadá, Chile, Estados Unidos e México, como sendo influenciadoras na arrecadação previdenciária. Nesse modelo, os coeficientes apresentaram influências diferentes para os países da América Latina, devido a diversos fatores socioeconômicos e locais.

Ademais, quando utilizados no ambiente da tarifação de seguros, segundo Lucubo (2021), essa determinação do preço do seguro é uma atividade fundamental no negócio de seguradoras. Em sua pesquisa, voltada às tarifas de seguros de automóveis, identifica três fatores de risco que compõem seu modelo: (1) fatores relacionados ao veículo, como valor, idade, categoria, marca e potência; (2) fatores relacionados ao condutor, tais como idade, gênero e tempo de habilitação; e (3) fatores relacionados à circulação, como zona de residência, quilometragem e uso do veículo. Aplicando um MLG, buscando um número mínimo de variáveis que melhor expliquem os dados, foram utilizados os métodos *backward* e *forward* para selecionar as variáveis independentes do modelo, que analisou 60.000 apólices de seguros automotivos, com 12 variáveis independentes. Ao fim do ajuste, apenas 7 das variáveis eram significativas, que foram capazes de prever a frequência de sinistralidade e o custo médio de cada sinistro.

Na análise de investimentos, os modelos de regressão linear também podem ser aplicados para, por exemplo, ajudar nas decisões sobre os investimentos de uma empresa petrolífera. Estudo desenvolvido por Bastos, Guimarães e Severo (2015), aplicou o MLG para justificar os investimentos da Petrobras de 2009 a 2013, que continha as variáveis independentes “valor de mercado” e “valor patrimonial”. Ainda, foi aplicado mais um modelo de regressão linear, analisando o ativo circulante e a correlação de suas contas, tendo como resultado, após o ajuste, de que esse ativo é explicado, principalmente, por caixa e equivalentes de caixa e contas a receber. Como resultados, foram definidos que, na análise dos investimentos,

é necessário realizar uma revisão da forma de investimento e, na análise do ativo circulante, fica evidente a relação positiva que se obtém maximizando o ativo circulante da empresa.

Esses modelos também podem ser aplicados na área da saúde, para prever a influência do peso, nível de HDL e o nível de triglicérides no colesterol de uma pessoa. Esse estudo foi desenvolvido por Rodrigues (2012), que projetou um modelo com as três variáveis. Pelo ajuste utilizando *stepwise* para selecionar variáveis independentes, verifica-se que o colesterol total é explicado pelo nível de triglicérides e de HDL. Também se notou que se trata de um modelo linear que segue uma distribuição normal de probabilidade, sendo esse caso um modelo linear múltiplo.

Outra possibilidade aplicada em modelos de regressão que não seguem uma distribuição normal de probabilidade é a normalização dos dados. Segundo Regis (2021), que aplicou um modelo de regressão linear múltipla para prever a mensalidade de escolas particulares no Brasil, a normalização aplicada dentro do *machine learning* [1]ajuda na obtenção de um melhor resultado do modelo. Como resultado do modelo ajustado e padronizado, ficou claro que o conjunto de variáveis foram influentes na predição desse valor de mensalidade, que apoiaram nas tomadas de decisões de custos de cada escola.

A análise estatística do futebol também usa modelos de regressão. A FIFA, Federação Internacional de Futebol, foi fundada em 1904 na Suíça, estabelecendo as regras do esporte coletivo mais jogado do mundo. Porém, foi apenas a partir da década de 1950 que a ciência começou a ser aplicada no futebol, como um meio de anotar passes e efetividade (Suhre, 2020)[2]. Não é segredo que esse esporte, assim como todos os outros, desenvolveu-se rapidamente no nível técnico e, cada vez mais disputado, os times buscam aplicar análises estatísticas para conseguir diversas vantagens específicas sobre seus concorrentes nas partidas.

Diversos outros estudos já foram realizados aplicando a regressão linear nas áreas relacionadas a esportes, em destaque para o futebol. Silva (2018) aplica um modelo para realizar a previsão do resultado do campeonato brasileiro de futebol de 2017, utilizando os resultados do ano anterior. Em seu estudo, 9 variáveis foram consideradas, sendo elas: média de chutes ao gol, gols por chute, gols sofridos, porcentagem dos desarmes certos, finalizações certas, passes certos, posse de bola, perda da posse por jogo e finalizações. Essas 9 variáveis foram utilizadas para explicar o número de pontos do Campeonato Brasileiro de 2017 e, por meio de um ajuste

[1] Machine learning: aprendizado de máquina, é uma área da inteligência artificial que treina computadores por algoritmos para identificar padrões e realizar previsões (Iberdrola, s.d).
[2] Disponível em: https://www.cienciadabola.com.br/blog/analise-desempenho-futebol

do modelo pelo procedimento *backward*, que resultou em apenas três variáveis como explicativas: gols por chute, gols sofridos e finalizações.

Outra análise que pode ser aplicada, de maneira mais individual no esporte, é o estudo da contribuição de cada posição dos 10 jogadores em uma partida, exceto os goleiros. Por meio de uma regressão logística binária, diversas variáveis foram empregadas para a previsão do resultado da partida, como vitória, empate ou derrota, analisando 5396 jogadores das 10 principais ligas de futebol. Cada liga apresentou resultados diferentes, mas, de modo geral, não houveram variáveis com alto grau de influência, fator que pode ser explicado pelo alto grau de ligação entre o resultado da partida e variáveis mais generalizadas, como o desempenho, a adaptação dos jogadores e a imprevisibilidade que permeia a partida como um todo. Esse estudo pode ser aplicado na prática não só como um indicador da contribuição das ações individuais de cada jogador em sua posição, mas também como uma ferramenta aplicada nos treinos visando a melhora da contribuição de posições estratégicas, de acordo com os próximos adversários e seus estilos de jogo. (Schimidt, 2012)

A atribuição de variáveis *dummy* em modelo de regressão é bastante comum, pois insere na análise variáveis qualitativas como forma binária. No estudo de Dantas et al (2009), que teve por objetivo verificar se existe relação entre a variação do preço de ações e medidas contábeis em entidades esportivas, inseriu a variável *dummy* como sendo a primeira colocação no campeonato italiano com valor 1, e 0 para qualquer outra colocação. Os dados, tratados pelo Excel, foram referentes às médias de preços das ações do time Juventus F.C, semestralmente, de 2002 a 2009. Foram aplicadas regressões simples, lineares e múltiplas, mas, os dois primeiros modelos aplicados, incluindo a variável *dummy*, não têm poder de previsão eficaz da variação dos preços das ações, sendo apenas o de regressão múltipla explicativo, apesar de não possuir um percentual de explicação satisfatório.

Em síntese, os modelos de regressão linear são aplicados à diversas áreas do conhecimento, para uma quantidade enorme de pesquisas para a previsão de seus resultados. O MLG pode ser mais acessível, de acordo com a base de dados, uma vez que sua distribuição de probabilidade pertence a uma família exponencial, não sendo, obrigatoriamente, uma distribuição normal.

3-FUNDAMENTAÇÃO TEÓRICA

3.1-MODELO DE REGRESSÃO LINEAR

A análise de regressão permite "encontrar alguma função que estime o comportamento do conjunto de dados que não se dispõe, a partir dos dados coletados" (Figueira, 2006). O termo "regressão" surgiu em 1885 com o estatístico Francis Galton em estudos de medidas do corpo humano, estendendo, atualmente, sua aplicação em diversas áreas do conhecimento.

De acordo com Flávia Chein (2019), a regressão linear pode ser definida como "um instrumento estatístico para, simplesmente, resumir dados e informações", ou seja, em sua análise, é verificada a dependência estatística entre as variáveis, podendo estimar a diferença entre um grupo de variáveis tratadas no modelo e um grupo de variáveis de controle dos resultados. No sentido mais técnico, Wooldridge (2018), define a regressão linear como a explicação de uma variável dependente $\boldsymbol{Y}$ em função de termos de $\boldsymbol{X}$, matriz de variáveis independentes.

O modelo geral de regressão linear, de acordo com as definições apresentadas por Charnet et al (2008), no livro "Análise de Modelos de Regressão Linear com Aplicações", diz respeito à relação linear entre a variável dependente Y e o aumento de uma unidade em X_j.

O modelo geral de regressão linear pode ser escrito como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Em que:

- y_i é a variável dependente
- x_{ij} é o valor da variável explicativa x_j
- β_k é o parâmetro que será estimado
- i = 1, ..., n
- j = 1, ..., p
- k = 0, ..., p

Além disso, o elemento ϵ_i, o erro do modelo, segue uma distribuição normal, com média zero e variância σ^2, ou seja, $\epsilon \sim N(0; \sigma^2)$.

Na forma matricial, o modelo de regressão linear pode ser representado pela equação abaixo:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}\text{, ou}$$

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ : \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ 1 & x_{31} & x_{32} & . & . & . & x_{3p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix} \text{ e } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}$$

Sendo:

- **Y** o vetor da variável dependente;
- **β** o vetor de coeficientes de dimensão *(p+1)* x *1*;
- **X** a matriz definida de acordo com o modelo, de dimensão *n* x *(p+1);*
- **Xβ** a função de regressão.

Nessa notação, o erro pode ser escrito como $\boldsymbol{\epsilon} \sim NM_n(\mathbf{0}; \sigma^2\boldsymbol{I})$, em que:

- **0** é o vetor nulo de dimensão *n x 1;*
- **I** é a matriz identidade de *n* x *n;*
- NM_n denota a distribuição normal multivariada de dimensão *n;*
- σ^2 é a variância do erro.

Sendo assim, sob um modelo de regressão linear $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, sendo $\boldsymbol{\epsilon} \sim NM_n(\mathbf{0};\ \sigma^2\boldsymbol{I})$, o $\boldsymbol{Y}$, correspondente aos valores de $\boldsymbol{x}$, também apresenta distribuição normal multivariada, de ordem *n*, com média, ou esperança, e matriz de variâncias e covariâncias dadas por:

$$E[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X\beta}$$

$$Var[\boldsymbol{Y}|\boldsymbol{X}] = \sigma^2\boldsymbol{I}$$

Logo, a distribuição de $\boldsymbol{Y}$ pode ser escrita como:

$$\boldsymbol{Y} \sim NM_n(\boldsymbol{X\beta}\ ;\ \sigma^2\boldsymbol{I})$$

3.2-O MÉTODO DOS MÍNIMOS QUADRADOS

O Método dos Mínimos Quadrados é utilizado para a estimação dos parâmetros $\boldsymbol{\beta}$. Nas palavras de Figueiredo Filho et al (2011), "dizer que o modelo é ajustado utilizando a forma funcional de mínimos quadrados ordinários significa que uma reta que minimiza a soma dos

quadrados dos resíduos será utilizada para resumir a relação linear entre *Y* e X_i". A utilidade e importância do método se encontram na minimização da soma dos quadrados dos erros do modelo, ou seja, maximiza a explicação que *X* tem sobre esse modelo.

De acordo com Charnet et al (2008), o método de quadrados mínimos pode ter sua aplicação estendida para modelo de regressão linear múltipla na forma matricial. Então, "o vetor de dimensão p + 1, cujos elementos compõem a solução de ajuste da função linear em β_0, $\beta_1, \beta_2, \beta_3, \dots, \beta_p$ a um conjunto de pontos $(y_1, x_{11}, x_{12}, x_{13}, \dots, x_{1p})$, ..., $(y_n, x_{n1}, x_{n2}, x_{n3}, \dots, x_{np})$ pelo método de quadrados mínimos, é dados por":

$$(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y},$$

desde que a inversa de $\boldsymbol{X'X}$ exista.

Dessa maneira, "o estimador de quadrados mínimos para o vetor de parâmetros $\boldsymbol{\beta}$ é definido conforme a solução de quadrados mínimos: "

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

Ademais, obtêm-se, simultaneamente, as esperanças, variâncias e covariâncias de cada elemento $\widehat{\boldsymbol{\beta}}$, utilizando a notação matricial:

$$E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$
$$Var[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X'X})^{-1}$$

Com isso, assim como a distribuição de probabilidade de $\boldsymbol{Y}$, a distribuição de $\widehat{\boldsymbol{\beta}}$ é normal $(p+1)$-variada, sendo que p + 1 é a dimensão de $\boldsymbol{\beta}$, ou seja:

$$\widehat{\boldsymbol{\beta}} \sim NM_{(p+1)}(\boldsymbol{\beta}\ ;\ \sigma^2(\boldsymbol{X'X})^{-1})$$

Além disso, é necessário encontrar um estimador não viciado para a variância do erro, σ^2, dado pela fórmula abaixo:

$$\widehat{\sigma^2} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n - p - 1}$$

De acordo com as hipóteses de Gauss-Markov, que serão apresentadas abaixo, tem-se que o estimador não viesado de σ^2 possui valor esperado de:

$$E\left(\widehat{\sigma^2}\right) = \sigma^2$$

De acordo com Wooldridge (2018), algumas hipóteses devem ser não rejeitadas, para que os estimadores encontrados sejam não viesados dos parâmetros da população.

1. O modelo é linear nos parâmetros, ou seja, a variável *y* está relacionada à variável independente *x* e ao erro ϵ, escrito por:

$$\mathrm{Y} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

2. A amostra de tamanho *n* deve ser aleatória, proveniente do modelo populacional acima.
3. A amostra, e a população, não possuem multicolinearidade perfeita, não existindo relacionamentos lineares perfeitos entre as variáveis independentes.
4. A média condicional do erro é zero:

$$E\left[\boldsymbol{\epsilon}\middle|x_1, x_2, \ldots, x_p\right] = E[\boldsymbol{\epsilon}] = 0$$

Que infere que $\widehat{\boldsymbol{\beta}}$ é não viesado para $\boldsymbol{\beta}$ se $E\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$.

5. Por fim, a quinta hipótese é a da homocedasticidade, que significa que o erro possui a mesma variância para qualquer valor da variável explicativa, ou seja, a variância é constante, dada por:

$$Var\left(\boldsymbol{\epsilon}|x_1, x_2, \ldots, x_p\right) = Var[\boldsymbol{\epsilon}] = \boldsymbol{\sigma^2}$$

Essas cinco hipóteses são conhecidas como as hipóteses de Gauss-Markov. Quando um modelo é submetido a essas cinco hipóteses, o estimador MQO $\widehat{\beta_j}$ para β_j é o melhor estimador linear não viesado. As três primeiras fazem alusão ao modelo, e as duas últimas são condições que os resíduos devem assumir e, a hipótese 5 é a que garante que o estimador é não viesado.

Ainda conforme Wooldridge (2018), com esse estimador, pode-se encontrar o Erro Padrão da Regressão, que é um estimador do desvio padrão do termo do erro, dado pela raiz quadrada positiva de σ^2. Para encontra-lo, pode-se utilizar a fórmula:

$$EP(\widehat{\boldsymbol{\beta}}) = \sqrt{Var(\widehat{\boldsymbol{\beta}})}$$

Essa fórmula só será válida se os erros exibirem homocedasticidade, ou seja, variância constante.

Posto isto, o Teorema de Gauss-Markov justifica o uso do método de MQO em vez de utilizar uma variedade de estimadores concorrentes. Esse teorema propõe que, sob as hipóteses 1 a 5, $\widehat{\beta_0}, \widehat{\beta_1}, \dots \widehat{\beta_p}$ são os melhores estimadores lineares não viesados de $\beta_0, \beta_1, \dots \beta_p$. Esses estimadores são os melhores, pois apresentam a menor variância. Uma importante aplicação desse teorema é que, com ele, não é necessário procurar por estimadores não viesados alternativos, pois aqueles obtidos por MQO são os melhores (Wooldridge, 2018).

Como os coeficientes betas têm distribuição normal, pode-se testar hipóteses para qualquer β_k com o objetivo de saber se o x_j associado a esse β_k, com j = k, contribui para explicar a variável resposta **Y**. Para isso, testam-se as seguintes hipóteses:

$$H_0: \beta_k = 0$$
$$H_1: \beta_k \neq 0$$

Para testá-las, utiliza-se a estatística *T*, em que os betas apresentam a seguinte distribuição:

$$T = \frac{\widehat{\beta_k} - \beta_k}{EP(\widehat{\beta_k})} \sim t_{(n-p-1)}, \text{ para } k = 0, 1, \dots, p.$$

Portanto, para um nível de significância α, a hipótese H_0 de que a contribuição da regressora não é significativa, deve ser rejeitada quando $|T_{\beta_k}| \geq t_{(\frac{\alpha}{2}, n-p-1)}$. Para os casos de testes unilaterais ($H_1: \beta_k > 0 \; ou \; H_1: \beta_k < 0$), H_0 será rejeitada se $|T_{\beta_k}| \geq t_{(\alpha, n-p-1)}$.

Observação: usa-se a distribuição *t~Student* pelo fato de não conhecermos a variância do modelo e, portanto, usamos uma estimativa, dada por $\hat{\sigma}^2$; caso σ^2 seja conhecido, usamos a distribuição normal para a estimação dos coeficientes betas.

Outro conceito importante é o teste de significância do modelo proposto, que será resumido em uma tabela ANOVA, do inglês, *Analysis of Variance*, que testa a significância do modelo. Em outras palavras, testa se existe algum x_j que explica, em média, parte da variabilidade de $\boldsymbol{Y}$. Para que este x_j contribua para explicar parte da variabilidade de $\boldsymbol{Y}$, o β_j associado a ele deve ser diferente de 0.

Para isso, se assumem as seguintes hipóteses:

$$H_0: \beta_1 = \beta_2 = \ ... = \ \beta_p = 0$$
$$H_1: pelo\ menos\ um\ \beta_j \neq 0$$

Se H_0 é rejeitada, então pode-se concluir que há uma contribuição significante de uma ou mais variáveis regressoras na explicação da variabilidade de **Y**. Para a composição da ANOVA do modelo de regressão linear múltipla, são utilizadas as somas de quadrados da regressão, *SQReg*, e dos resíduos, *SQE*, para formar a soma de quadrados total de **Y**, *SQT*. Em termos de variabilidade, o modelo de regressão linear decompõe a variabilidade total do **Y** em variabilidade devido à regressão e variabilidade devido ao erro. Assim, entende-se por variabilidade a soma de quadrados:

$$SQT = SQReg + SQE$$

Compostos por:

- $SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \boldsymbol{Y'Y} - n\bar{y}^2$
- $SQE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \boldsymbol{Y'Y} - \boldsymbol{\hat{\beta}X'Y}$
- $SQReg = SQT - SQE = \boldsymbol{\hat{\beta}X'Y} - n\bar{y}^2$

Cada uma dessas relações pode ser explicada de maneira teórica, sendo *SQT* a variação total de **Y** em torno de sua média, obtido pelo quadrado da diferença entre os valores observados e a média; *SQReg* a variação das esperanças específicas de **Y** em torno da sua média, obtido pelo quadrado da diferença entre os valores estimados e a média; e *SQE* a variação de **Y** em torno da reta, obtido pelo quadrado da diferença entre os valores observados e os valores estimados. Com isso, tem-se a tabela de ANOVA:

Tabela 1- ANOVA

FV	GL	SQ	QM	F
Regressão	p	SQReg	$QMReg = \frac{SQReg}{p}$	$\frac{SQReg/p}{SQE/(n-p-1)}$
Erro	$n-p-1$	SQE	$QME = \frac{SQE}{n-p-1} = \sigma^2$	
Total	$n-1$	SQT		

Fonte: elaboração própria

Em que:

- FV significa fonte de variação
- GL são os graus de liberdade
- SQ são as somas de quadrados
- QM são os quadrados médios: QM$Reg \sim \chi^2_p$ e QME$\sim \chi^2_{n-p-1}$
- F é a estatística de Fisher: $F = \frac{QMReg}{QME} \sim F_{p,n-p-1}$

Para que o modelo seja significativo, a estatística F precisa assumir valores suficientemente altos para que garanta a rejeição da hipótese nula. Para que isto ocorra, a SQReg deve ser suficientemente maior que SQE. Em outras palavras, se $F > F_{p,n-p-1}(\alpha)$, então rejeita-se a H_0 ao nível α de significância. Na prática, rejeitar essa H_0 quer dizer que existe modelo que explica parte da variabilidade de **Y**, com ao menos um x_j.

Outra análise importante, dentro da multicolinearidade, é a observação do fator de inflação da variância (VIF). Os fatores de VIF são os elementos da diagonal da matriz $(X^T X)^{-1}$, e representam o incremento da variância devido à presença de multicolinearidade. Seu valor pode ser calculado pela fórmula abaixo:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Em que:

- R_j^2 é o coeficiente de correlação múltipla
- $j: 1, \dots, k$

De acordo com Salvian apud Montgomery, Peck, Vining (2016), "se o valor de R_j^2 for próximo a 1, isto significa que existe uma alta correlação entre a variável X_j e as demais variáveis, então $1 - R_j^2$ estará próximo de 0 e consequentemente, o VIF assumirá um valor grande". Com isso, se o valor do VIF for maior que 10, indica que a multicolinearidade pode influenciar as estimativas de MQO.

Por fim, na parte final da avaliação do modelo, define-se o coeficiente de determinação, o R^2, que indica a proporção da variabilidade total da variável dependente, explicada estatisticamente por uma ou mais variáveis independentes. Esse coeficiente é geralmente expresso em porcentagens, e é dado pela fórmula:

$$R^2_{Y|X_1,X_2,\dots,X_p} = \frac{SQReg}{SQT} = 1 - \frac{SQE}{SQT},$$

em que $0 \leq R^2 \leq 1$. Além disso, é importante dar ênfase ao R^2 ajustado, que tenta corrigir a superestimação do R^2, já que ele aumenta à medida que são incluídas variáveis, que pode ser calculado pela seguinte fórmula:

$$R^2 ajustado = 1 - \frac{QME}{QMT}$$

3.3-MÉTODOS DE SELEÇÃO DE VARIÁVEIS INDEPENDENTES

Além de diversos pressupostos e ajustes que devem ser considerados em um modelo de regressão linear, é necessário também focar a atenção nas variáveis independentes que vão compor o modelo. Segundo Charnet et al (2008), "devemos decidir se incluir todas as variáveis regressoras disponíveis ou incluir apenas um subconjunto dessas variáveis". Após essa decisão, verificam-se a significância e a adequação desse modelo ajustado, além dos resíduos. Ribon et al (2014) ressalta que, para a escolha das variáveis que compõem a equação que melhor irão explicar essa influência das variáveis independentes, há alguns métodos comumente utilizados, dentre eles os chamados "*backward, forward*" e "*stepwise*".

A utilização do "backward", ou "passo atrás", consiste

> Na obtenção de um modelo, onde inicia-se pela equação na qual todas as variáveis estão inclusas e estas são eliminadas uma por vez a cada passo a partir de um nível de significância pré-estabelecido, ou seja, as variáveis eliminadas vão apresentar o nível de significância mais elevado, dentre as variáveis incluídas, até que o nível de significância do valor de F de todas as variáveis seja menor que o do modelo pré-estabelecido. (RIBON et al, 2014, p.31).

De acordo com os autores, no modelo completo dessa etapa, são investigadas as contribuições individuais das variáveis no modelo, eliminando a que apresentar o pior desempenho no critério mínimo exigido. Depois, compara-se o modelo completo com o modelo reduzido, pela retirada dessa variável, pelas somas de quadrados. A estatística do teste da contribuição dessa variável é dada por:

$$\frac{SQReg^c - SQReg^r}{\widehat{\sigma^2}}$$

Onde:

- $SQReg^c$ é a soma dos quadrados da regressão do modelo completo
- $SQReg^r$ é a soma dos quadrados da regressão do modelo reduzido

A contribuição da variável é significante se o valor dessa estatística for maior que um quantil especificado da distribuição de F com 1 e ($n - m - 1$) graus de liberdade, sendo m o número de

variáveis do modelo completo. Mesmo que o modelo apresente várias variáveis insignificantes, é retirada aquela que apresenta a menor significância de todas, uma por vez. Esse processo é seguido até que todas as variáveis apresentem significância maior que a estatística. Resumidamente, Charnet et al (2008) apresenta 4 etapas a serem seguidas no modelo:

1. Ajustar o modelo completo de *m* variáveis e obter $SQReg^c$ e $\widehat{\sigma^2}$;
2. Para cada uma das *m* variáveis do modelo completo do passo 1, considerar o modelo reduzido (retirando essa variável) e calcula $SQReg^r$ para obter o valor da estatística apresentada acima;
3. Achar o mínimo dos *m* valores da estatística obtidos no passo 2, denotado por F_{min};
4. Seja F_{out} o quantil especificado da distribuição F com 1 e ($n - m - 1$) graus de liberdade. Se $F_{min} > F_{out}$, o processo é interrompido e opta pelo modelo completo dessa etapa, e se $F_{min} < F_{out}$, voltar ao passo 1, iniciando uma nova etapa em que o modelo completo tem ($m - 1$) variáveis, pois se eliminou a variável cuja estatística era igual a F_{min}.

   Já na seleção "forward", ou "passo a frente"

   > O procedimento se inicia com nenhuma outra variável na equação e se adiciona uma por vez, iniciando-se pela variável que promover um decréscimo mais significativo na soma de quadrados do resíduo e que apresente um nível de significância do valor de F, menor que um valor pré-estabelecido, até que todas as variáveis com nível de significância do valor de F menor do que esse valor estejam dentro do modelo. (RIBON et al, 2014, p.31-32).

Assim, se em uma etapa não houver a inclusão de nenhuma variável, o processo é interrompido, definindo o modelo final. Em uma dada etapa, com um determinado modelo, realiza-se uma comparação com modelos que incluem uma nova variável, escolhendo o que apresentar melhor desempenho no critério mínimo exigido. Resumindo também em quatro passos esse método (Charnet et al, 2008), tem-se:

1. Ajustar o modelo reduzido de *m* variáveis e obter $SQReg^r$;
2. Para cada variável que não pertença ao modelo do passo 1, considerar o modelo completo com a adição desta variável extra e calcular $SQReg^c$ e $\widehat{\sigma^2}$ para obter o valor da estatística;
3. Achar o máximo dos valores dessa estatística obtidos no passo 2, denotado por F_{max};
4. Seja F_{in} o quantil especificado da distribuição F com 1 e ($n - m - 2$) graus de liberdade. Se $F_{max} > F_{in}$, voltar ao passo 1, iniciando uma nova etapa onde o modelo reduzido tem ($m + 1$) variáveis, pois inclui a variável com estatística igual a F_{max}, e se $F_{max} < F_{in}$, interrompe-se o processo e escolhe o modelo reduzido dessa etapa.

Por fim, o procedimento “stepwise”, ou “passo a passo”, torna o próprio procedimento completo, pois

> Todas as variáveis incluídas no modelo apresentam nível de significância do valor de F menor que o valor pré-estabelecido para eliminação de variáveis, e todas as variáveis fora do modelo têm nível de significância do valor de F maior que o valor pré-estabelecido para inclusão de variáveis. (RIBON et al, 2014, p.32).

Este procedimento é uma generalização do “passo à frente”, já que após cada etapa de incorporação de uma variável, há uma etapa em que uma das variáveis já adicionadas pode vir a ser descartada. O procedimento chega ao fim quando nenhuma variável é incluída ou retirada do modelo analisado. Sendo assim, o método pode ser resumido em oito passos, também segundo Charnet et al (2008), descritos abaixo.

1. Ajustar o modelo reduzido de *m* variáveis e obter $SQReg^r$;
2. Para cada variável que não pertença ao modelo do passo 1, considerar o modelo completo com a adição desta variável extra e calcular $SQReg^c$ e $\widehat{\sigma^2}$ para obter o valor da estatística;
3. Achar o máximo dos valores dessa estatística obtidos no passo 2, denotado por F_{max};
4. Seja F_{in} o quantil especificado da distribuição F com 1 e ($n - m - 2$) graus de liberdade. Se $F_{max} > F_{in}$, passar ao passo 5, com modelo completo composto por (m + 1) variáveis, sendo elas as *m* variáveis do modelo do passo 1 e a variável cuja estatística é igual a F_{max}, e se $F_{max} < F_{in}$, passar ao passo 5 com modelo completo igual ao modelo do passo 1 ou encerrar o processo se, no passo 8 da etapa anterior, nenhuma variável tiver sido eliminada;
5. Ajustar o modelo completo de *k* variáveis, sendo *k* igual a *m* ou (m + 1), e obter $SQReg^c$ e $\widehat{\sigma^2}$.
6. Para cada uma das *k* variáveis do modelo completo do passo 5, considerar um modelo reduzido, retirando esta variável, e calcular $SQReg^r$;
7. Achar o mínimo dos *k* valores da estatística obtidos no passo 6, denotado por F_{min};
8. Seja F_{out} o quantil especificado da distribuição F com 1 e (n - k – 1) graus de liberdade. Se $F_{min} > F_{out}$, não eliminar nenhuma variável e voltar ao passo 1, iniciando uma nova etapa com modelo reduzido com *k* variáveis ou encerrar o processo se no passo 4 nenhuma variável for anexada, e se $F_{min} < F_{out}$, eliminar a variável com estatística igual a F_{min} e voltar ao passo 1, iniciando uma nova etapa com modelo reduzido com (k – 1) variáveis

Ainda, dentro desses métodos, de forma computadorizada, é possível aplicar o Critério de Informação de Akaike, ou AIC, que nada mais é do que, também, um critério de seleção de variáveis independentes de um modelo (Cavalaro, 2019). Sua fórmula, descrita abaixo, busca selecionar o modelo com o menor valor de AIC. Portanto, retira-se do modelo a variável com o p-valor maior que 0,15 e observa-se o valor de AIC. Segue-se esse passo até que não se possam mais retirar variáveis e o AIC do modelo ajustado seja menor que o do modelo anterior.

$$AIC = -2\log\left(L\left(\hat{\theta}|dados\right)\right) - 2k$$

Sendo:

- $\log\left(L\left(\hat{\theta}|dados\right)\right) - k$, em que $L\left(\hat{\theta}|dados\right)$ é a função de verossimilhança.

3.4-ANÁLISE DE RESÍDUOS

Em concordância com as análises acima, ainda, para verificar se um modelo é adequado, é necessário analisar seus resíduos. Para Morettin (2017), a análise de resíduos é estudar o comportamento do modelo pelo conjunto de dados observados, por meio das discrepâncias entre os valores observados e os valores ajustados pelo modelo. Desse modo, o que se faz é analisar "o comportamento individual e conjunto destes resíduos, comparando com as suposições feitas sobre os verdadeiros erros ϵ_i" (Morettin, 2017).

Como já descrito nas seções 3.1 e 3.2, temos:

$$\boldsymbol{Y} \sim NM_n(\boldsymbol{X\beta};\ \sigma^2\boldsymbol{I})$$

$$\widehat{\boldsymbol{\beta}} \sim NM_p(\boldsymbol{\beta}; \sigma^2(\boldsymbol{X'X})^{-1})$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{X\beta};\ \sigma^2\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'})$$

Posto isto, tem-se a esperança e variância dos resíduos, dadas por:

$$E[\boldsymbol{\epsilon}] = \boldsymbol{0}$$

$$Var[\boldsymbol{\epsilon}] = \sigma^2[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}]$$

Portanto, os resíduos apresentam a distribuição:

$$\boldsymbol{\epsilon} \sim NM_n(\boldsymbol{0};\ \sigma^2[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}]$$

Como posto, os resíduos apresentam uma distribuição normal. Além disso, a consistência é um requisito mínimo de um estimador. Nas palavras de Wooldridge (2018), "seja $\widehat{\beta_J}$ o estimador de MQO de β_j para algum j. Para cada n, $\widehat{\beta_J}$ tem uma distribuição de probabilidade. Como $\widehat{\beta_J}$ é não viesado sob as 4 hipóteses de Gauss-Markov, essa distribuição

tem valor médio β_j. Se esse estimador for consistente, a distribuição de $\widehat{\beta_j}$ se torna mais e mais estreitamente distribuída ao redor de β_j quando o tamanho da amostra cresce." Em outras palavras, o estimador se aproxima do valor populacional quando a amostra cresce. Com isso, como o valor esperado do erro é 0, implica que qualquer função das variáveis explicativas é não correlacionada com ϵ, ou seja, a correlação entre as variáveis independentes e os resíduos é 0.

Assim como a média condicional do modelo deve ser diferente de 0, a correlação entre ϵ e qualquer uma das variáveis $x_1, x_2, \dots, x_k$ também causam viés dos estimadores de mínimos quadrados ordinários, fazendo com que todos os estimadores sejam inconsistentes. Essa inconsistência em $\widehat{\beta_j}$ é chamada de viés assimptótico. Quando o tamanho da amostra aumenta, não é possível testar as hipóteses sobre os parâmetros, sendo necessário que os parâmetros sigam uma distribuição normal, decorrente do erro apresentar essa mesma distribuição de probabilidade. Essa normalidade do erro é garantida pela sexta hipótese que se pode testar em um modelo de regressão linear múltipla: o erro populacional ϵ é independente das variáveis explicativas x_i e é normalmente distribuído com média 0 e variância σ^2, ou seja, $\epsilon \sim N(0; \sigma^2)$.

Essa hipótese da normalidade não afeta na existência de viés nos estimadores e nem que o método MQO produz os melhores estimadores, mas a inferência exata baseada nas estatísticas t e F necessita de que os resíduos sigam uma distribuição normal de probabilidade. Uma alternativa é se apoiar no Teorema do Limite Central, que conclui que os estimadores de mínimos quadrados ordinários satisfazem a normalidade assimptótica, ou seja, são aproximadamente normalmente distribuídos em amostras de tamanhos suficientemente grandes. Em suma:

- $\sqrt{n}(\widehat{\beta_j} - \beta_j) \sim$ Normal(0;σ^2/a_j^2), em que $\sigma^2/a_j^2 > 0$ é a variância assimptótica de $\sqrt{n}(\widehat{\beta_j} - \beta_j)$
- $\widehat{\sigma^2}$ é um estimador consistente de σ^2=Var(ϵ)
- Para cada j, $\frac{\widehat{\beta_j} - \beta_j}{dp(\widehat{\beta_j})} \sim N(0,1)$ e $\frac{\widehat{\beta_j} - \beta_j}{ep(\widehat{\beta_j})} \sim N(0,1)$

Este teorema é útil, visto que, se os y_i não sejam provenientes de uma distribuição normal, podemos usá-lo para concluir que os estimadores de MQO satisfazem a normalidade assimptótica para amostras suficientemente grandes, presumindo média condicional 0 e homocedasticidade do erro (Wooldridge, 2018).

Para analisar os resíduos de um modelo, são observadas as ocorrências de três hipóteses principais: (*i*) independência do erro; (*ii*) normalidade e (*iii*) variância constante.

Para testar a independência, dois testes são comumente utilizados: o teste Qui-Quadrado, utilizado para descobrir se existe uma associação entre a variável da linha e a da coluna, e o teste de Durbin-Watson, que testa a presença de autocorrelação entre os resíduos do modelo. Segundo Gujarati (2011), a estatística *d* pode ser definida como:

$$d = \frac{\sum_{i=2}^{n}(\widehat{\epsilon_i} - \widehat{\epsilon_{i-1}})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$

Esse teste se baseia nos resíduos estimados, e tem diversos valores tabulados pelos cientistas que o desenvolveram. Se o valor de *d* estiver fora dessas zonas críticas, é possível a verificação da autocorrelação. Com H_0: ausência de autocorrelação, a hipótese é não rejeitada se o valor de *d* for próximo de 2 e, com H_1: presença de autocorrelação, é não rejeitada se os valores são muito distantes de 2. Também é possível a análise desse teste pelo seu p-valor, em que, se o valor de p for maior que α, nível de significância, a H_0 é não rejeitada, e os resíduos são independentes.

A segunda hipótese, de que os erros seguem uma distribuição normal de probabilidade, pode ser testada pelo teste de Kolmogorov-Smirnov e pelo teste de Shapiro-Wilk. Neste trabalho, será utilizado o teste de Shapiro-Wilk para a normalidade dos resíduos. Com H_0: os erros têm distribuição normal e H_1: os erros não têm distribuição normal (Anjos, 2005), e pode ser calculado por:

$$W = \frac{b^2}{s^2} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{(\sum_{i=1}^{n}(y_i - \bar{y_i}))^2}$$

Onde:

- y_i é a variável aleatória observada;
- a_i são coeficientes tabelados;
- s^2 é a soma dos quadrados dos erros

Para a avaliação desse teste, é observado o p-valor resultante, de forma computacional. Se o p-valor for maior que α, então a hipótese nula é não rejeitada, e os resíduos apresentam distribuição normal.

Por fim, o teste da homocedasticidade pode ser realizado pelos testes de Breusch-Pagan ou Goldfeld-Quandt. O teste de Breusch-Pagan, tem por hipóteses H_0: há homocedasticidade do erro, ou seja, Var(ϵ)=σ^2, e H_1: há heterocedasticidade, podendo ser verificadas pelo teste BP, que é a versão LM do teste da heterocedasticidade. Ou seja, considerando como hipótese nula que a hipótese 5 da seção 3.2 é verdadeira, então $H_0: Var(\epsilon|x_1, x_2, \dots, x_k) = \sigma^2$ e, como é suposto que ϵ tem média condicional 0, então a hipótese nula de homocedasticidade pode ser escrita como $H_0: E(\epsilon^2|x_1, x_2, \dots, x_k) = E(\epsilon^2) = \sigma^2$. Isso mostra que, para testar a violação da

hipótese de homocedasticidade, verifica-se se ϵ^2 está relacionado, em média, a uma ou mais variáveis explicativas. Segundo Wooldridge (2018), esse teste pode ser resumido em 3 passos:

1. Avaliar o modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$ por MQO e obter os quadrados dos resíduos, $\hat{\epsilon}^2$;
2. Executar a regressão de $\hat{\epsilon}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + erro$, em que δ_k são as variáveis explicativas do erro, e anotar o R^2 dessa regressão;
3. Construir a estatística LM e calcular o p-valor da distribuição χ_k^2 e, se o p-valor for suficientemente pequeno, abaixo do nível de significância selecionado, a H_0 é rejeitada, sendo necessário realizar algum ajuste.

3.5-O MODELO DE PROBABILIDADE LINEAR - MPL

Segundo Wooldridge (2018), para explicar um resultado qualitativo utilizando uma regressão múltipla, pode-se aplicar um modelo de probabilidade linear que resulta em um resultado 0 ou 1, ou seja, a variável dependente $\boldsymbol{Y}$ é binária e assume somente os valores 0 ou 1. Como $\boldsymbol{Y}$ assume somente esses dois valores, β_j não pode ser interpretado como a mudança em $\boldsymbol{Y}$ em razão do aumento de uma unidade em x_j. Ou seja, a variável dependente só muda de 0 para 1 ou de 1 para 0, ou, simplesmente, não muda.

As hipóteses de Gauss-Markov são válidas e, no caso da probabilidade condicional, a probabilidade de sucesso pode ser escrita como:

$$E[\boldsymbol{Y}|\boldsymbol{X}] = P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

em que $\mathbf{X}$ representa todas as variáveis independentes. Ainda, $P(y = 1|\mathbf{X})$ pode ser chamado de probabilidade de resposta.

O modelo de probabilidade linear é chamado assim porque a probabilidade de resposta é linear nos parâmetros β_j, que medem a mudança na probabilidade de sucesso quando x_j muda, mantendo fixos os outros fatores:

$$\Delta P(y = 1|\boldsymbol{X}) = \beta_j \Delta x_j$$

Nesse modelo, o método dos mínimos quadrados ordinários segue o mesmo procedimento caso fosse um modelo múltiplo. Um passo importante da interpretação do MPL

é a definição do que constitui o “sucesso” no modelo, e é mais viável dar um nome que descreva o evento $y = 1$ à variável dependente. Apesar de ser um modelo de fácil interpretação, possui algumas deficiências. A primeira a ser destacada é que, se são agregadas certas combinações de valores das variáveis independentes, pode-se obter previsões menores que 0 ou maiores que 1. A segunda é que a probabilidade não pode ser linearmente relacionada às variáveis independentes em todos os seus possíveis valores. Mesmo com esses dois problemas, o MPL é útil e frequentemente utilizado, e funciona bem com os valores das variáveis independentes que estejam próximos das médias da amostra.

Além disso, por ter natureza binária, a variância de $\boldsymbol{Y}$ condicional em $\mathbf{X}$ infringe a hipótese de Gauss-Markov, ou seja:

$$Var(\boldsymbol{Y}|\boldsymbol{X}) = p(\boldsymbol{X})[1 - p(\boldsymbol{X})]$$

Neste caso, sendo $p(\boldsymbol{x})$ a probabilidade de sucesso. Posto isto, fica claro que pode haver heterocedasticidade no modelo de probabilidade linear, que não causa viés nos estimadores de mínimos quadrados ordinários, mas afeta as estatísticas F e T.

4-APLICAÇÃO DE UM MODELO DE PROBABILIDADE LINEAR

4.1-MATERIAIS E MÉTODOS

Nesta seção, será abordada uma aplicação de um modelo de probabilidade linear para explicar a probabilidade do Palmeiras vencer um jogo do Campeonato Brasileiro de 2023. Foram coletados os dados de 33 rodadas do Campeonato Brasileiro, série A, de 2023, para o Palmeiras, no site FootStats. Os dados são referentes apenas a 33 rodadas, pois o campeonato ainda está em andamento durante a construção do modelo. O modelo proposto explica a vitória desse time, variável *dummy* que recebe 1 em caso de vitória e 0 se não ocorrer vitória, em função de 4 variáveis: Mando do jogo, Finalizações por jogo, Porcentagem da posse de bola por jogo, e Número de gols por jogo. Dessas 4 variáveis, Mando de jogo é uma variável dummy, sendo atribuída 0 se o Palmeiras não é o time mandante e 1 se é o time mandante.

O método aplicado para explicar essa probabilidade é um modelo de regressão de probabilidade linear.

Para avaliação do modelo, serão aplicados testes apresentados na fundamentação teórica deste trabalho, utilizando o software estatístico R (R Core Team, 2023). O modelo foi construído, primeiramente, com as 4 variáveis, para apresentar a probabilidade de vitória do Palmeiras. Abaixo, segue um quadro que especifica as variáveis utilizadas:

Quadro 1-Variáveis

Variável	**Sigla utilizada**	**Composição**
Dependente: Vitória	Vitória	Variável dummy, sendo atribuído 1 quando ocorre vitória, e 0 quando não ocorre.
Independente: Mando de jogo	Mando	Variável dummy, sendo 1 atribuído ao mando de jogo, e 0 quando o mando é do adversário
Independente: Finalizações por jogo	Finalizações	Contagem das finalizações por jogo
Independente: Porcentagem da posse de bola	Posse	Contagem da porcentagem da posse de bola por jogo
Independente: Número de gols por jogo	Gols	Contagem dos gols feitos por jogo

Fonte: elaboração própria

Com isso, o modelo completo pode ser escrito como:

$$E[Vitória|\boldsymbol{X}] = \beta_0 + \beta_1 Mando + \beta_2 Finalizações + \beta_3 Posse + \beta_4 Gols,$$

em que **X** = (Mando, Finalizações, Posse, Gols)

4.2- RESULTADOS E DISCUSSÃO

O modelo foi estimado por MQO e a tabela com os coeficientes estimados segue abaixo:

Tabela 2-Resumo estatístico do modelo completo

Coeficientes	**Valor estimado**	**Erro padrão**	***p*-valor de *t***
β_0	0,099046	0,560703	0,8610
β_1	0,273507	0,145736	0,0710
β_2	0,01158	0,019686	0,5610
β_3	-0,004397	0,012378	0,7250
β_4	0,213447	0,045879	0,0000

Fonte: elaboração própria

Por esses resultados, o modelo pode ser escrito da forma:

$$E[Vitória|\boldsymbol{X}] = 0{,}099 + 0{,}273Mando + 0{,}011Finalizações - 0{,}004Posse + 0{,}213Gols$$

O AIC do modelo é 35,4153. Como existe a presença de variáveis que não são significativas, é necessário realizar um ajuste do modelo. Neste trabalho, será utilizado o método *backward* para a seleção das variáveis independentes, com AIC para selecionar modelo, que já foi citado na seção 3.3. Retirando a variável Posse, que apresenta o maior p-valor dentre os betas, o novo modelo passa a ser escrito como:

$$E[Vitória|\boldsymbol{X}] = \beta_0 + \beta_1 Mando + \beta_2 Finalizações + \beta_3 Gols$$

O novo modelo também foi estimado por MQO, apresentando os resultados abaixo:

Tabela 3-Resumo estatístico do modelo retirando a variável Posse

Coeficientes	**Valor estimado**	**Erro padrão**	***p*-valor de *t***
β_0	-0,081135	0,235373	0,7328
β_1	0,269877	0,143170	0,0695
β_2	0,007857	0,016412	0,6357
β_3	0,215824	0,044699	0,0000

Fonte: elaboração própria

Ou seja, pode ser escrito como;

$$E[Vitória|\boldsymbol{X}] = -0{,}0811 + 0{,}2698Mando + 0{,}0078Finalizações + 0{,}2158Gols$$

O AIC desse novo modelo é menor que o anterior, com valor de 33,56365. O modelo ainda apresenta variáveis que não são explicativas, e será aplicado o método backward novamente. A variável retirada agora é Finalizações, pois apresenta o maior p-valor, no valor de 0,6357, e o modelo passa a apresentar as seguintes informações:

Tabela 4-Resumo estatístico do modelo retirando a variável Finalizações

Coeficientes	**Valor estimado**	**Erro padrão**	***p*-valor de *t***
β_0	0,01794	0,11068	0,8723
β_1	0,29993	0,12701	0,0249
β_2	0,21904	0,04362	0,0000

Fonte: elaboração própria

Agora, não é mais necessário retirar variáveis do modelo, pois os p-valores dos novos betas estimados são todos significativos, exceto para o β_0, ou seja, menores que 0,05. Além disso, o AIC desse modelo também é o menor de todos os 3, com valor de 31,8243, finalizando o método de escolha das variáveis independentes. Logo, o modelo final pode ser escrito como:

$$E[Vitória|\boldsymbol{X}] = 0{,}018 + 0{,}30Mando + 0{,}22Gols$$

Esse modelo apresenta um R^2 igual a 0,4853, sendo o melhor dos 3 modelos testados, explicando, em média, 48,53% da variabilidade total do Y. Pode-se falar também que, em média, quando o Palmeiras tem o mando de campo, a probabilidade de vencer aumenta em 0,30 ou 30% e também que, quando o time faz um gol no jogo, a probabilidade de vitória aumenta, em média, 0,22 ou 22%.

Por fim, é necessário realizar os testes sobre os resíduos do modelo, descritos na seção 3.4. O primeiro teste a ser aplicado é o da independência dos resíduos, realizado pelo teste de Durbin-Watson, que resultou em um p-valor de 0,572, não rejeitando H_0, ao nível de 5% de significância, ou seja, não há autocorrelação e os resíduos são independentes.

O segundo teste a ser aplicado é o da homocedasticidade, realizado por meio do teste de Breusch-Pagan. Este teste resultou em um p-valor de 0,2297, ou seja, H_0 é não rejeitada, e os resíduos são homocedásticos, apresentando variação constante.

O terceiro teste que será aplicado, e o último, é o da normalidade dos resíduos, feito pelo teste de Shapiro-Wilk e, como resultou em um p-valor de 0,027, a H_0 é rejeitada ao nível de 5% de significância, ou seja, os resíduos não apresentam distribuição normal de probabilidade e, consequentemente, Y também não segue uma distribuição normal. Porém, é possível se apoiar no Teorema do Limite Central, que conclui que os estimadores de mínimos quadrados ordinários satisfazem a normalidade assimptótica, ou seja, são aproximadamente normalmente distribuídos em amostras de tamanhos suficientemente grandes. Logo, será assumido que os resíduos são aproximadamente distribuídos normalmente, já que os estimadores de MQO satisfazem a normalidade assimptótica para amostras suficientemente grandes, presumindo média condicional 0 e homocedasticidade do erro, testes aceitos acima.

5-CONSIDERAÇÕES FINAIS

Tendo em vista que a aplicação de um modelo de probabilidade linear pode ser estendida a diversos contextos, não apenas para realizar a previsão de vitória de um time de futebol, isto mostra como os modelos de regressão, não apenas os MPL, podem ser aplicados a diversos estudos, desde áreas das ciências sociais aplicadas até áreas da saúde e de esportes.

Apesar de toda essa variedade de aplicações, muitas das vezes as variáveis que pela literatura podem ser influências nos resultados, na prática podem não apresentar significância, sendo necessário realizar uma seleção de variáveis regressoras no modelo, por métodos como o *backward*, *forward* ou *stepwise*, além das observações do AIC. Além disso, a análise dos resíduos do modelo ajustado também se faz necessária, seguindo pressupostos de normalidade, independência e homocedasticidade pois, além de mostrarem que o modelo necessita de um ajuste, também influenciam para que não haja viés nos estimadores de mínimos quadrados ordinários e se o modelo final encontrado é adequado para uso.

Ainda, sobre os modelos com variáveis *dummies*, são úteis pelo fato de incorporarem essas variáveis, que abordam resultados categóricos, incluídos como variáveis resposta ou regressoras binárias no modelo, para abordarem características como sexo, região, acesso a determinado recurso, religião, setor de trabalho, se uma pessoa contrata um plano de saúde ou não e, no exemplo aplicado, se o Palmeiras é o time mandante ou não. Essas variáveis se mostram de importante aplicação pois incorporam aos modelos resultados influentes em políticas públicas e outras análises sócio econômicas.

Dentro do mundo dos esportes, em destaque para o futebol, as aplicações de estudos estatísticos e suas abordagens é recente, datando de 1950, como uma anotação de passes e efetividade. A partir disso, a estatística no futebol só cresceu, de forma mais computadorizada e detalhada, abordando aspectos como número de gols, time mandante, público, distância do gol, número de finalizações, porcentagem de aproveitamento de cada jogador e de cada posição, dentre outros, como formas de prever resultados tanto das partidas como de campeonatos disputados por diversos times (Mangerona, 2023).

A aplicação proposta neste trabalho aborda os conceitos apresentados na fundamentação teórica, se mostrando útil, com um bom ajuste, após aplicações de seleção de variáveis regressoras e análise de resíduos. O modelo final consegue explicar 48,53% da variabilidade total da variável independente. Na análise dos resíduos, estes seguem os pressupostos da independência e da homocedasticidade a um nível de significância de 5%, mas não são

exatamente normalmente distribuídos, sendo considerada a normalidade assimptótica, uma vez que os estimadores de MQO satisfazem essa normalidade, presumindo média condicional 0 e homocedasticidade. Além disso, em média, quando o Palmeiras tem o mando de campo, a probabilidade de vencer aumenta em 0,30 ou 30% e também que, quando o time faz um gol no jogo, a probabilidade de vitória aumenta, em média, 0,22 ou 22%.

Cabe destacar que este trabalho apresenta limitações, tanto teóricas quanto práticas, mas que podem ser diminuídas ou até corrigidas, estendendo o modelo de probabilidade linear para um modelo com mais variáveis, além da possível aplicação de outros tipos de modelos lineares generalizados. Portanto, o modelo encontrado aqui pode ser melhorado, utilizando mais variáveis que possam prever a vitória de um time, além da construção de um modelo que possa abranger mais times de uma vez, atualizando as estatísticas conforme o passar do tempo.

Como estudo futuro, pretende-se trabalhar com novas aplicações de MLGs para analisar dados sobre futebol, usando modelos que possam analisar não só a probabilidade de vitória de um time, mas o lucro do time, pontos obtidos em algum campeonato, probabilidade de ser campeão, dentre outras.

REFERÊNCIAS

ANJOS, A. **Teste de Shapiro-Wilk para Normalidade**, 2005.

AOKI, J. **Abordagens Matemáticas e Estatística para o Futebol**. Trabalho de Conclusão de Curso, Universidade Estadual de Campinas, 2010.

BASTOS, E.; GUIMARÃES, J.; SEVERO, E. Modelo de Regressão Linear para Análise de Investimentos em uma Empresa do Ramo Petrolífero. **Revista de Produção e Desenvolvimento**, v.01, n.1, p.77-88, 2015.

BUSSAB, W.; MORETTIN, P. **Estatística Básica**. 9^0 Edição. São Paulo: Saraiva, 2017.

CAVALARO, L. **Um Procedimento para Seleção de Variáveis em Modelos Lineares Generalizados Duplos**. Dissertação de Mestrado (Programa Interinstitucional de Pós-Graduação em Estatística), Universidade Federal de São Carlos e Universidade de São Paulo, 2019.

CAVALCANTE, J.; REINALDO, L. Modelos Lineares Generalizados (MLG'S) e sua Aplicação em Ciências Atuariais In: VIII SIMPÓSIO DE ATUÁRIA. **Anais...**, Fortaleza, 2019.

CHARNET, R. et al. **Análise de Modelos de Regressão Linear com Aplicações**. 2^0 Edição. Campinas: Editora Unicamp, 2008.

CHEIN, F. **Introdução aos Modelos de Regressão Linear**. Brasília: Enap, 2019.

DANTAS, M. et al. O comportamento do preço das ações de clubes de futebol mediante a variação de aspectos contábeis: o estudo de caso do Juventus F.C. – Itália. **Revista Ambiente Contábil**, v.1, n.2, p.55-67, 2009.

DAVID, L. **Modelos Lineares Generalizados Multinomiais Univariado e Bivariado**. Trabalho de Conclusão de Curso (Departamento de Estatística), Universidade Federal de São Carlos, 2021.

FIGUEIRA, C. **Modelos de Regressão Logística**. Dissertação de mestrado (Programa de Pós-Graduação em Matemática), Universidade Federal do Rio Grande do Sul, 2006.

FIGUEIREDO FILHO, D. et al. O que fazer e o que não fazer com a Regressão: Pressupostos e Aplicações do Modelo Linear de Mínimos Quadrados Ordinários (MQO). **Revista Política Hoje**, v.20, n.1, 2011.

GOMES, A. Introdução à Variável Dummy. **Blog de Estatística da Prof. Fernanda Maciel**, s.d. Disponível em: https://blog.proffernandamaciel.com.br/variavel_dummy/ Acesso em 23 de novembro de 2023.

GUJARATI, D.; PORTER, D. **Econometria Básica**. 5^0 Edição. São Paulo: AMGH Editora Ltda., 2011.

IBERDROLA. O que é Machine Learning. Iberdrola, s.d. Disponível em: https://www.iberdrola.com/inovacao/o-que-e-machinelearning#:~:text=Machine%20Learning%20%C3%A9%20uma%20disciplina,fazer%20previs%C3%B5es%20(an%C3%A1lise%20preditiva) Acesso em: 23 de novembro de 2023.

LUCUBO, A. **Seguro de Responsabilidade Civil Automóvel Modelos de Tarifação**. Dissertação de Mestrado (Actuariado, Estatística e Investigação Operacional), Universidade Nova de Lisboa, 2020.

MANGERONA, R. **A Estatística no Futebol:** Uma Análise dos Principais Fatores que Influenciam o Número de Gols Feitos pelos Jogadores no Campeonato Inglês. Trabalho de Conclusão de Curso (Departamento de Estatística), Universidade Federal de São Carlos, 2023.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to linearregression analysis. **John, Wiley and Sons**, Inc., New York, 612p, 2006.

NELDER, J.; WEDDERBURN, R. Generalized linear models. **Journal of the Royal Statistical Society** A 135, 370-384, 1972.

OLIVEIRA FILHO, M. A Utilização da Regressão Linear como Ferramenta Estratégica para a Projeção dos Custos de Produção In: IX CONGRESSO BRASILEIRO DE CUSTOS. **Anais**..., São Paulo, 2002.

PINTO, L. et al. Análise dos Fatores que Influenciam a Arrecadação do Regime Geral em Países do Continente Americano. **Revista Pensamento e Realidade**, v.34, n.2, p.137-155, 2019.

REGIS, R. **Modelo de Regressão Linear para Predição de Mensalidades de Escolas Particulares do Brasil**. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Produção), Universidade Federal do Estado do Rio de Janeiro, 2021.

RIBON, A. et al. Seleção de Método Estatístico para Modelos de Estimativa da Qualidade Física de Solos Argilosos. **Revista de Agricultura**, v.89, n.1, p.29-40, 2014.

RODRIGUES, S. **Modelo de Regressão Linear e suas Aplicações**. Tese de Doutorado, Universidade da Beira Interior, 2012.

SANTOS, V. Regressão à Média: A Tendência à Mediocridade. FM2S-Educação e Consultoria, 2016. Disponível em: https://www.fm2s.com.br/public/blog/regressao-media-tendencia-mediocridade/amp Acesso em: 23 de novembro de 2023.

SALVIAN, M. Multicolinearidade. XV Congresso USP de Iniciação Científica em Contabilidade. **Anais...**, Piracicaba, 2016.

SCHIMIDT, V. **Quais Ações são significativas para uma Vitória no Futebol? Uma Análise por Regressão Logística**. Trabalho de Conclusão de Curso (Instituto de Biociências), Universidade Estadual Paulista, 2021.

SILVA, B. Regressão Linear Múltipla Aplicada ao Futebol. **Revista Brasileira de Futsal e Futebol**, v.10, n.38, p.262-270, 2018.

SUHRE, C. Os Dados e a Análise de Desempenho no Futebol. **Ciência da Bola**, 2023. Disponível em: https://www.cienciadabola.com.br/blog/analise-desempenho-futebol Acesso em: 23 de novembro de 2023.

WOOLDRIDGE, J. **Introdução à Econometria**. 3^0 Edição. São Paulo: Cengage Learning, 2018.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria.<https://www.R-project.org/>